Docket No. 50269-0569 *Patent*

UNITED STATES PATENT APPLICATION

FOR

SUGGESTING AN ALTERNATIVE TO THE SPELLING OF A SEARCH QUERY

INVENTORS:

ARKADY BORKOVSKY
CHAD CARSON
MARCO ZAGHA

PREPARED BY:

HICKMAN PALERMO TRUONG & BECKER LLP
1600 WILLOW STREET
SAN JOSE, CALIFORNIA 95125
(408) 414-1080

"Express Mail" mailing label number EV323351480US

Date of Deposit February 27, 2004

SUGGESTING AN ALTERNATIVE TO THE SPELLING OF A SEARCH QUERY

CLAIM OF PRIORITY

[0001] The present application claims priority to and is a continuation-in-part of U.S. Patent Application Serial No. 10/364,078, entitled "SUGGESTING AN ALTERNATIVE TO THE SPELLING OF A SEARCH QUERY", filed February 10, 2003; which claims priority to U.S. Provisional Patent Application No. 60/411,535, entitled "SPELLING IN WEB SEARCH", filed September 17, 2002, and U.S. Provisional Patent Application No. 60/413,092, entitled "SPELLING IN WEB SEARCH", filed September 23, 2002. The entire contents of the foregoing patent applications are hereby incorporated by reference.

FIELD OF THE INVENTION

[0002] The present invention relates to search engines, and in particular, to suggesting an alternative spelling for a search query.

BACKGROUND OF THE INVENTION

[0003] A search engine is a computer program that helps a user to locate information. Using a search engine, a user can enter one or more search query terms and obtain a list of resources that contain or are associated with subject matter that matches those search query terms. While search engines may be applied in a variety of contexts, search engines are especially useful for locating resources that are accessible through the Internet. Resources that may be located through a search engine include, for example, files whose content is composed in a page description language such as Hypertext Markup Language (HTML). Such files are typically called pages. One can use a search engine to generate a list of

Universal Resource Locators (URLs) and/or HTML links to files, or pages, that are likely to be of interest.

[0004] Some search engines order a list of files before presenting the list to a user. To order a list of files, a search engine may assign a rank to each file in the list. When the list is sorted by rank, a file with a relatively higher rank may be placed closer to the head of the list than a file with a relatively lower rank. The user, when presented with the sorted list, sees the most highly ranked files first. To aid the user in his search, a search engine may rank the files according to relevance. Relevance is a measure of how closely the subject matter of the file matches query terms.

[0005] To find the most relevant files, search engines typically try to select, from among a plurality of files, files that include many or all of the words that a user entered into a search request. Unfortunately, the files in which a user may be most interested are too often files that do not literally include the words that the user entered into the search request. If the user has misspelled a word in the search request, then the search engine may fail to select files in which the correctly spelled word occurs.

[0006] Worse yet, a user may enter, into a search request, a word that is a correctly spelled word, but that is not the word that means the thing for which the user desires to search. For example, a user who wants to find files that include information about "Silicon Valley" may, through ignorance or by accident, request a search for "Silicone Valley". Because "Silicone" is a correctly spelled word, a spelling checking program will not detect any error. Under such circumstances, the user is likely to obtain a list of results that have little to do with what the user was actually looking for.

[0007] A user may successfully enter a search request that includes correctly spelled words that are used in the correct context. Even in this case, a search engine may fail to return many existing files that include information in which the user would be very interested. Search results may be under-inclusive for a variety of reasons. A verb in the search request may be in a different verb tense than the verbs contained in the files. A noun in the search request may be expressed in the plural form while the nouns in the files are expressed in the singular form. A word may have more than one correct spelling, and the spelling used in the files might be different than the spelling that the user selected. The words included in the files may be synonyms of the words that the user entered into the search request. For any of these or other reasons, a search engine may return sub-optimal results.

[0008] The approaches described in this section are approaches that could be pursued, but not necessarily approaches that have been previously conceived or pursued. Therefore, unless otherwise indicated, it should not be assumed that any of the approaches described in this section qualify as prior art merely by virtue of their inclusion in this section.

SUMMARY OF THE INVENTION

[0009] Techniques are provided for suggesting an alternative spelling for a search query. According to one embodiment, in response to a user entering a spelling in a search query, a search engine selects, from a plurality of files, one or more files that contain one or more spellings that are similar to, but not exactly the same as, the user-entered spelling. The files may be selected because the files contain the user-entered spelling, or because the files are linked to by links, such as HTML links, which contain the user-entered spelling. Thus, the selected files do not need to contain the user-entered spelling in order to be selected. The search engine searches the selected files for spellings that are similar to the user-entered spelling. In searching the files for similar spellings, the search engine may search the entirety of the files, rather than only those portions of the files that would be displayed as result fields (e.g., a title, an abstract, a URL, etc.). As a result, the search engine locates one or more instances of spellings that are similar to, but not exactly the same as, the user-entered spelling. In response to locating the similar spellings, the search engine adds the similar spellings to a list of candidate alternative spellings, which may be filtered and presented to a user.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] The present invention is illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

[0011] Figure 1 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for suggesting an alternative spelling to a search query;

[0012] Figure 2 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative spelling based on the frequencies of occurrences of two different spellings in a plurality of files;

[0013] Figure 3 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative spelling based on the frequencies of occurrences of files, in a plurality of files, that contain one of two different spellings;

[0014] Figure 4 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative spelling based on the frequencies of occurrences of two different spellings in one or more previously received search requests;

[0015] Figure 5 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative spelling based on the frequencies of occurrences of files, in a plurality of files, which contain at least two of at least three different words or phrases;

[0016] Figure 6 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative spelling based on the frequencies of occurrences of previously received search requests that contain at least two of at least three different words or phrases;

[0017] Figure 7 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for filtering a set of candidate alternative spellings based on the frequencies with which those candidate alternative spellings were accepted in the past;

[0018] Figure 8 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for determining, based on links to a file, whether a spelling should be added to a set of possible alternative spellings;

[0019] Figures 9A and 9B are flow diagrams that illustrate a technique, according to an embodiment of the present invention, for determining whether to retain or remove a first candidate alternative spelling from a set of candidate alternative spellings based on how frequently the first candidate alternative spelling occurs in result fields obtained from a search for a second candidate alternative spelling; and

[0020] Figure 10 is a block diagram that illustrates a computer system 1000 upon which an embodiment of the invention may be implemented.

DETAILED DESCRIPTION OF THE INVENTION

[0021] A method and apparatus for suggesting an alternative spelling to a search query is described. In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present invention.

FUNCTIONAL OVERVIEW

[0022] It is contemplated that this technique, and all of the techniques described herein, may be applied to sequences of words (i.e., phrases) as well as to individual words. Figure 1 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for suggesting an alternative spelling to a search query. In block 102, a request is received. The request is to search for files that match a search query. The request may be to search for files that contain several specific words. The plurality of files that are considered during the search may include one or more pages that are accessible through the Internet. For example, an Internet search engine may receive, from a user, a request to search for pages that contain the word "Silicon".

[0023] In block 104, an alternative spelling is selected. As used herein, "spelling" refers to a particular spelling of one or more words. Thus, two different spellings may be two completely different words or sequences of words, or two different words or sequences of words that are spelled similarly to, but not exactly the same as, each other. As used herein, a "word" may be either a word that occurs in a natural language, such as English, or a special notation, such as a product code, URL, or ISBN. For example, the word "http:" is often misspelled as "htpp:". The alternative spelling is selected based on a set of factors. The set of factors may include how frequently the alternative spelling occurs in the plurality of files. The set of factors may include how frequently the alternative spelling occurs in one or more previously received requests to search for files. The alternative spelling may be similar to the corresponding part of the original query for which the alternative spelling is selected. For example, the alternative spelling may be spelled similarly to the corresponding part of the original query. For another example, the alternative spelling may be pronounced similarly to the corresponding part of the original query. For yet another example, the alternative spelling may be synonymous to the corresponding part of the original query. The alternative spelling may be selected from among a set of candidate alternative spellings. A different alternative spelling may be selected for each different corresponding part of a search query. While in one embodiment, only one alternative spelling is selected for each corresponding part of the original query, in an alternative embodiment, more than one alternative spelling may be selected for each corresponding part of the original query. Furthermore, multiple words may be selected as an alternative spelling of a single word, and vice-versa. In other words, words in spellings may be split or joined. For example, the spelling "New York" may be selected as an alternative to the spelling "Newyork."

[0024] For example, the search engine may submit the search query to a spelling checking routine. In response, the spelling checking routine may consult a stored dictionary file of spellings and return one or more alternative spellings that are spelled similarly to, but not exactly the same as, the submitted search query. The search engine may then determine how frequently one or more words from each alternative spelling occurs in the Internet pages. In making this determination, the search engine may consult a pre-generated index of spellings that indicates, for each spelling, how frequently that spelling occurs in a set of

Internet pages that has been analyzed for spelling frequency. The more frequently an alternative spelling occurs in the set of Internet pages, the more likely it is that the alternative spelling would produce search results in accord with the interests of the user who requested the search.

[0025] The search engine may also, or instead, determine how frequently each alternative spelling occurs in one or more previously received requests to search the set of Internet pages. In making this determination, the search engine may consult a log of search requests that the search engine has previously received. The log may include each spelling of each previously entered search query. Some spellings may occur more frequently in the log than other spellings. Alternatively, the search engine may consult an index that indicates, for each spelling that occurs in the log, how frequently that spelling occurs in the log. The more frequently an alternative spelling occurs in the log, the more likely it is that the alternative spelling would produce search results in accord with the interests of the user who requested the search.

[0026] While in one embodiment, both of the factors described above are used to select an alternative spelling, in alternative embodiments, only one or the other of these factors is used to select an alternative spelling.

[0027] Before consulting a full index of spellings to determine the frequency of an occurrence of a spelling in a plurality of files or previously received search requests, a search engine may consult a cache. The cache may contain a subset of spellings (with associated frequencies) from the full index that are most often searched for, or most recently searched for, in the full index. If a spelling is not contained in the cache, then the search engine may consult the full index.

[0028] In block 106, the alternative spelling is optionally presented to a user. For example, the search engine may present, to the user who entered the search request, a page that asks the user if he meant to enter the alternative spelling instead of the spelling that he entered. The search engine may additionally accept input from the user that indicates whether the user prefers to use the alternative spelling instead of the spelling that the user entered.

[0029] While in one embodiment, the alternative spelling is presented to a user, in an alternative embodiment, the alternative spelling is automatically, without any input from the user, substituted for the spelling that the user originally entered in the search query. Thus, in various embodiments, a user may or may not be permitted to choose whether the alternative spelling will be substituted for the originally entered spelling in the search query.

[0030] While in one embodiment, the alternative spelling is presented to the user before the search engine performs a search for the originally entered spelling, in an alternative embodiment, the alternative spelling is presented to the user after the search engine performs a search for the originally entered spelling.

[0031] While in one embodiment, only one alternative spelling is presented to the user, in an alternative embodiment, more than one spelling may be presented to the user. While in one embodiment, an alternative spelling is always presented to the user, in an alternative embodiment, an alternative spelling is only presented to the user under specified circumstances.

[0032] Thus, a search engine may aid a user in obtaining search results that are of the most interest to the user by suggesting alternative spellings that are gleaned from various sources and evaluated according to various criteria. With each search request, the sources from which alternative spellings are gleaned may grow and become more refined. Viewed from one perspective, the search engine is capable of learning. Consequently, the search engine may more intelligently select an alternative spelling to suggest to a user.

SELECTING THE ALTERNATIVE SPELLING BASED ON HOW FREQUENTLY THE ALTERNATIVE SPELLING OCCURS IN THE PLURALITY OF FILES

[0033] A candidate alternative spelling may be evaluated for selection based on the frequency of that candidate alternative spelling relative to the frequency of a particular spelling for which the candidate alternative spelling is an alternative. This helps to avoid the suggestion of an alternative spelling that, when used instead of another spelling in a search request, would actually cause less interesting results to be provided to a user than if the other spelling had been used.

[0034] Figure 2 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative spelling based on the frequencies of occurrences of two different spellings in a plurality of files. In block 202, it is determined how frequently a particular spelling occurs in a plurality of files. For example, a search engine may consult a pre-generated index, such as is discussed above, to determine how frequently a particular spelling, which was entered as part of a search request, occurs in a set of pages.

[0035] In block 204, it is determined how frequently an alternative spelling occurs in the plurality of files. For example, a search engine may consult a pre-generated index to determine how frequently an alternative spelling occurs in a set of pages. The alternative spelling may be a spelling that was suggested by a spelling checking routine as being a possible alternative spelling of the particular spelling that was entered in the search request.

[0036] In block 206, based on a set of one or more rules that weigh the frequency of occurrences of the particular spelling against the frequency of occurrences of the alternative spelling, it is determined whether to present the alternative spelling to a user. For example, a search engine may apply a rule that states that the alternative spelling should not be presented to a user if the frequency of occurrences of the particular spelling is greater than the frequency of occurrences of the alternative spelling.

[0037] A search engine administrator may construct such a rule and store the rule in a set of rules that the search engine applies when evaluating candidate alternative spellings for presentation to a user. Rules may take the form of conditional statements. The variety of different rules that could be constructed is practically limitless.

[0038] Figure 3 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative spelling based on the frequencies of occurrences of files, in a plurality of files, that contain one of two different spellings. While the technique described above with reference to Figure 2 evaluates an alternative spelling based on a total spelling frequency, the technique described below evaluates an alternative spelling based on a frequency of occurrences of files that contain one or more instances of the alternative spelling. The technique described below may be used to give less weight to alternative spellings that occur many times in each of only a relatively few pages, and more weight to alternative spellings that occur only a relatively few times in each of many pages.

[0039] In block 302, it is determined how frequently files that contain a particular spelling occur in a plurality of files. For example, a search engine may consult a pre-generated index to determine how frequently pages that include a particular spelling, which was entered as part of a search request, occur in a set of pages.

[0040] In block 304, it is determined how frequently files that contain an alternative spelling occur in the plurality of files. For example, a search engine may consult a pre-generated index to determine how frequently pages that include an alternative spelling occur in a set of pages.

[0041] In block 306, based on a set of one or more rules that weigh the frequency of occurrences of files that contain the particular spelling against the frequency of occurrences of files that contain the alternative spelling, it is determined whether to present the alternative spelling to a user. For example, a search engine may apply a rule that states that the alternative spelling should not be presented to a user if the frequency of occurrences of files that contain the particular spelling is greater than the frequency of occurrences of files that contain the alternative spelling.

[0042] While in one embodiment, the determination whether to present an alternative spelling is based on a frequency of occurrences of that alternative spelling in a plurality of files, in an alternative embodiment, that determination is based on a relevance score of one or more pages that contain the alternative spelling. Relevance scores may be determined by a variety of techniques. For example, a page's relevance score may be based on how many words, from the search request, are contained in the page, and/or how frequently those words occur in the page. A page's relevance score may also take into account some "page quality metrics" like a citation index, the authoritativeness of the page source, etc. A page's relevance may also be defined relative to the user who submitted the query. For example, relevance may be based on the language used to submit the query (e.g., English or French). For another example, relevance may be based on properties of the user (e.g., the user's geographic region, gender, or social group). Thus, relevance scores for pages that result from query submitted by a user in the United Kingdom may be determined differently than relevance scores for pages that result from the same query submitted by a user in the United States. A search engine may compare the relevance scores of one or more pages that would result from a search request that contains the particular spelling with relevance scores of one or more pages that would result from a search request that contains the alternative spelling instead of the particular spelling.

SELECTING THE ALTERNATIVE SPELLING BASED ON HOW FREQUENTLY THE ALTERNATIVE SPELLING OCCURS IN THE PREVIOUSLY RECEIVED SEARCH REQUESTS

[0043] A search engine may receive thousands of search requests. The search engine may maintain a log of the spellings used in these search requests. Some identical or nearly identical search requests may be submitted to a search engine multiple times. The frequency with which a given spelling is contained in prior search requests may be a strong indicator of how likely that spelling was meant to be used instead of a different spelling.

[0044] Figure 4 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative spelling based on the frequencies of occurrences of two different spellings in one or more previously received search requests. In block 402, it is determined how frequently a particular spelling occurs in one or more previously received search requests. For example, a search engine may consult a pre-generated index to determine how frequently a particular spelling, which was entered as part of the search request, occurs in a set of search requests that the search engine received previously to the search request. Such an index may be adjusted after each new search request.

[0045] In block 404, it is determined how frequently an alternative spelling occurs in the previously received search requests. For example, a search engine may consult a pre-generated index to determine how frequently an alternative spelling occurs in the previously received search requests.

[0046] In block 406, based on a set of one or more rules that weigh the frequency of occurrences of the particular spelling against the frequency of occurrences of the alternative spelling, it is determined whether to present the alternative spelling to a user. For example, a search engine may apply a rule that states that the alternative spelling should not be presented to a user if the frequency of occurrences of the particular spelling is greater than the frequency of occurrences of the alternative spelling.

[0047] In some embodiments, the determination whether to present the alternative spelling to a user is based on how the alternative spelling compares to other alternative spellings, and not merely based on how the alternative spelling compares to the original spelling.

CONTEXT-SENSITIVE ALTERNATIVE SPELLING SELECTION

[0048] When a user enters more than one word into a search request, one or more of those words may be used in the wrong context relative to the other words. For example, a user who wants to search for information relating to "Silicon Valley" may mistakenly submit a search request for "Silicone Valley." The frequency with which a given spelling occurs in the same file, or in the same previously received search request, as another word may be a good indicator of whether the given spelling was used in a correct context in the search request that includes that other word. For example, one would expect that a search engine would encounter more files and search requests that contain both the words "Silicon" and "Valley" than files and search requests that contain both the words "Silicone" and "Valley." Thus, a search engine may confidently suggest to a user who entered "Silicone Valley" that the user might have wished to enter the alternative "Silicon Valley" instead.

[0049] Figure 5 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative spelling based on the frequencies of occurrences of files, in a plurality of files, which contain at least two of at least three different words or phrases. In block 502, it is determined how frequently files, in which each file contains a particular spelling and a second word or phrase, occur in a plurality of files. For example, a search engine may determine how frequently pages, which contain both "Silicone" and "Valley", occur in a set of pages.

[0050] While in one embodiment, the second word or phrase is a word or phrase that was entered in a search request with the particular spelling, in an alternative embodiment, the second word or phrase may be obtained from a statistical repository that contains information on the frequencies with which certain words or phrases appear in the same pages as other words or phrases.

[0051] In block 504, it is determined how frequently files, in which each file contains an alternative spelling and the second word or phrase, occur in the plurality of files. For example, a search engine may determine how frequently pages, which contain both "Silicon" and "Valley", occur in a set of pages.

[0052] In block 506, based on a set of rules that weigh the frequency of occurrences of files, in which each file contains both the particular spelling and the second word or phrase, against the frequency of occurrences of files, in which each file contains the alternative spelling and the second word or phrase, it is determined whether to present the alternative spelling to a user. For example, a search engine may apply a rule that states that the alternative spelling should not be presented to a user if the frequency of occurrences of pages that contain both the particular spelling and the second word or phrase is greater than the frequency of occurrences of pages that contain both the alternative spelling and the second word or phrase. In other words, the rule may cause a search engine to not present an alternative spelling "Silicon" if "Silicone" occurs in pages with "Valley" more frequently than "Silicon" occurs in pages with "Valley".

[0053] While in one embodiment, the determinations described above determine the frequencies with which a particular spelling or alternative spelling occurs in the same files as a second word or phrase, in an alternative embodiment, the determinations determine the frequencies with which a particular spelling or alternative spelling occurs within some proximity of (e.g., immediately next to, within some distance of words as, in the same sentence as, in the same paragraph as, etc.) the second word or phrase.

[0054] While in one embodiment, a frequency of pages that include both the second word or phrase and either the particular spelling or alternative spelling is determined, in an alternative embodiment, a frequency of pages that include both a set of words or phrases and either the particular spelling or the alternative spelling is determined.

[0055] Figure 6 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for selecting an alternative word based on the frequencies of occurrences of previously received search requests which contain at least two of at least three different words or phrases. In block 602, it is determined how frequently previously received search requests that contain both a particular spelling and a second word or phrase occur in one or more previously received search requests. For example, a search engine may determine how frequently previously received search requests, which contain both "Silicone" and "Valley", occur in a set of previously received search requests.

[0056] While in one embodiment, the second word or phrase is a word or phrase that was entered in a search request with the particular spelling, in an alternative embodiment, the second word or phrase may be obtained from a statistical repository that contains information on the frequencies with which certain words or phrases appear in the same previously received search requests as other words or phrases.

[0057] In block 604, it is determined how frequently previously received search requests that contain both an alternative spelling and the second word or phrase occur in the previously received search requests. For example, a search engine may determine how frequently previously received search requests, which contain both "Silicon" and "Valley", occur in a set of previously received search requests.

[0058] In block 606, based on a set of rules that weigh the frequency of occurrences of previously received search requests that contain both the particular spelling and the second word or phrase against the frequency of occurrences of previously received search requests that contain both the alternative spelling and the second word or phrase, it is determined whether to present the alternative spelling to a user. For example, a search engine may apply a rule that states that the alternative spelling should not be presented to a user if the frequency of occurrences of previously received search requests that contain both the particular spelling and the second word or phrase is greater than the frequency of occurrences of previously received search requests that contain both the alternative spelling and the second word or phrase. In other words, the rule may cause a search engine to not present an alternative spelling "Silicon" if "Silicone" occurs in previously received search requests with "Valley" more frequently than "Silicon" occurs in previously received search requests with "Valley".

SELECTING THE ALTERNATIVE SPELLING BASED ON HOW FREQUENTLY THE ALTERNATIVE SPELLING WAS ACCEPTED

[0059] As described above, an alternative spelling may be presented to a user as an alternative to a spelling that the user entered in a search request. The user may have the option of accepting or rejecting a suggested alternative spelling. When the user either accepts or rejects an alternative spelling, that acceptance or rejection may be recorded for future use. A record may be made of the acceptance or rejection of the alternative spelling relative to the particular spelling to which the alternative spelling was suggested as an alternative, and/or relative to one or more of the other words or phrases that were entered in the search request. The frequency with which an alternative spelling is accepted or rejected may be a good indicator of whether the alternative spelling should be suggested under similar circumstances in the future.

[0060] Figure 7 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for filtering a set of candidate alternative spellings based on the frequencies with which those candidate alternative spellings were accepted in the past. In block 702, a set of candidate alternative spellings is established. For example, the set of candidate alternative spellings may include spellings that were selected by a spelling checking routine and/or a routine that selects synonyms of received spellings. The set of candidate alternative spellings may include spellings that were determined, based on a set of rules and other determinations such as those described above, to be alternative spellings that could be presented to a user.

[0061] In block 704, the set of candidate spellings is filtered based on a frequency with which each candidate spelling was accepted. For example, a search engine may remove, from a list of candidate alternative spellings, any candidate alternative spellings that are associated with an accept/reject ratio that is less than a specified threshold (e.g., 50%). While in one embodiment, candidate spellings are filtered based on acceptance rate even if those candidate spellings have never presented to a user, in an alternative embodiment, only those candidate spellings that have actually been accepted or rejected by a user are filtered based on acceptance rate.

[0062] An acceptance rate associated with a candidate alternative spelling may be a total acceptance rate of the candidate alternative spelling, an acceptance rate of the candidate alternative spelling relative to a particular word or phrase, or an acceptance rate of the candidate alternative spelling relative to a particular set of words or phrases.

[0063] In one embodiment, after the set of candidate alternative spellings has been filtered, the spellings remaining in the set of candidate alternative spellings may be presented to a user as suggested alternatives to a particular spelling that the user entered in a search request.

SELECTING THE ALTERNATIVE SPELLING BASED ON RULES

[0064] As described above, one or more rules may be specified by which candidate alternative spellings may be evaluated in determining whether those candidate alternative spellings should be suggested to a user. Such rules may include specified thresholds. If a threshold that is associated with a rule is met or exceeded by a particular number (e.g., score, frequency, etc.) that is associated with a candidate alternative spelling, then the candidate alternative spelling satisfies the rule. If a number that is associated with a candidate alternative spelling does not meet a threshold that is associated with a rule, then the candidate alternative spelling does not satisfy the rule.

[0065] In one embodiment, different weights may be associated with different sources from which candidate alternative spellings are gleaned. For example, a rule may indicate that, at least under specified circumstances, the frequency of occurrence in previously received search requests should be a greater factor in determining whether to present an alternative spelling to a user than the frequency of occurrence in a plurality of files.

[0066] While in one embodiment, candidate alternative spellings that do not satisfy one or more specified rules are not presented to a user, in an alternative embodiment, a candidate alternative spelling may be presented to a user even if that candidate alternative spelling does not satisfy one or more specified rules. For example, according to one embodiment, in one out of every N alternative spelling suggestions, a suggestion may include an alternative spelling that does not satisfy one or more specified rules. The frequency with which a suggestion may include an alternative spelling that does not satisfy one or more rules may itself be specified. An alternative spelling that is suggested in spite of not satisfying one or more rules may be selected based on how close the alternative spelling came to satisfying one or more rules (i.e., how close a number associated with the alternative spelling was to one or more thresholds specified by one or more rules).

[0067] Thus, in one embodiment, a user may potentially accept a candidate alternative spelling even if that candidate alternative spelling did not satisfy some specified rule, especially if the candidate alternative spelling came very close to satisfying the specified rule. The acceptance of a candidate alternative spelling that did not satisfy some rule may be an indicator that the unsatisfied rule should be modified.

EXTENDING THE SET OF POSSIBLE ALTERNATIVE SPELLINGS

[0068] As described above, a candidate alternative spelling may initially be selected from some repository of spellings, such as a dictionary file or a thesaurus file. Such a repository may be viewed as a set of possible alternative spellings. Once a candidate alternative spelling is selected from the repository (e.g., due to that alternative spelling being spelled similarly to, but not exactly the same as, a particular spelling that a user entered into a search request), that candidate alternative spelling may be further evaluated in a determination whether to ultimately present that candidate alternative spelling to a user.

[0069] In one embodiment, the set of possible alternative spellings (i.e., the spelling repository, dictionary, thesaurus, etc.) may be extended to include spellings that were not originally contained in the set. For example, a proper noun, such as a person's name, which is unlikely to be contained in a standard dictionary, but may be contained in several pages, may be added to a dictionary file upon the satisfaction of some criteria.

[0070] Spellings that are included in links to files, such as the spellings that are contained in the label of a hypertext link to a page, may be more likely to be spelled correctly and used in a proper context than spellings that are not contained in links. Spellings contained in links may be referred to as "anchor text". These spellings may be especially suitable for addition in the set of possible alternative spellings.

[0071] Figure 8 is a flow diagram that illustrates a technique, according to an embodiment of the present invention, for determining, based on links to a file, whether a spelling should be added to a set of possible alternative spellings. The technique is especially applicable for determining which version of a spelling should be added to a dictionary when two similar but unequal spellings are used in different links to a same page.

[0072] In block 802, a first number of links, which (a) link to a particular file and (b) include a first spelling, is determined. For example, to determine such a number, a search engine may consult a pre-generated index that contains, for each page and each spelling that appears in a link to that page, a number of times that a spelling occurs.

[0073] In block 804, a second number of links, which (a) also link to the particular file and (b) include a second spelling, is determined. The second spelling may be, for example, a spelling that is similar to, but not exactly the same as, the first spelling.

[0074] In block 806, it is determined whether the first number of links is greater than the second number of links. If the first number of links is greater than the second number of links, then, in block 808, the first spelling is added to a set of candidate alternative spellings. If the second number of links is greater than the first number of links, then, in block 810, the second spelling is added to the set of candidate alternative spellings. The set of candidate alternative spellings may be a dictionary file that is consulted by a spelling checking routine.

[0075] For example, a search engine could receive a search request for "Abraham Lincon". The search engine might not find either the spelling "Lincoln" or the spelling "Lincon" in a dictionary file. The search engine may determine that the spelling "Lincoln" occurs in ten links that link to a particular page, and that the spelling "Lincon" occurs in only one link to the same particular page. Based on this determination, the search engine may confidently add the spelling "Lincoln" to the dictionary file and omit the spelling "Lincon" from the dictionary file. It is likely, given the evidence provided in the anchor text, that the spelling "Lincon" is a misspelling that should not be added to the dictionary file.

[0076] In one embodiment, the file to which the links refer is also inspected to determine if the first spelling or the second spelling is contained in the file itself. In one embodiment, the frequency with which a spelling occurs in a file itself is given more weight than the frequency with which a spelling occurs in a link to that file. Therefore, even if the spelling "Lincon" occurs in more links to a page than the spelling "Lincoln", the search engine would add the spelling "Lincoln", and not the spelling "Lincon", to the dictionary file if the spelling "Lincoln" occurred more frequently than the spelling "Lincon" within the page itself.

GENERATING AND FILTERING CANDIDATE ALTERNATIVE SPELLINGS BASED ON HOW FREQUENTLY SPELLINGS OCCUR IN RESULT FIELDS

[0077] Result fields are fields that are typically displayed for each search result in a list of displayed search results. Search results typically include, for each result, a title of a page, one or more abstracts of a page, and a Uniform Resource Locator (URL) of a page. In one embodiment, candidate alternative spellings are generated based on how frequently certain spellings occur in result fields. In one embodiment, a set of candidate alternative spellings is filtered based on how frequently those candidate alternative spellings occur in result fields.

[0078] By looking for matches in the text of results fields, the asymmetry between anchor text and page content may be leveraged in order to generate and filter candidate alternative spellings. Anchor text may contain more misspellings than text that is contained in a title of a page, an abstract of a page, or a URL.

[0079] In one embodiment, to generate alternative candidate spellings, a page is searched for two spellings that are similar to, but not exactly the same as, each other. Then, for each spelling, a search is performed for files, in a plurality of files, which contain that spelling. Thus, for each spelling, a list of search results is obtained. Each result typically includes one or more result fields.

[0080] Then, for each of the two sets of results (one for each spelling), one or more of the result fields for each result is searched for the other one of the two spellings (i.e., the spelling that was not used to generate the results). A frequency of each spelling in the result fields generated by a search for the other spelling is obtained. That is, a frequency of a first spelling in the result fields generated by a search for a second spelling is obtained, and a frequency of the second spelling in the result fields of the first spelling is obtained. Based on a comparison between the two frequencies, one spelling may be selected over the other for inclusion in a list of candidate alternative spellings. In one embodiment, only a specified number of results are searched (e.g., the top ten results, ranked by relevance, for each spelling).

[0081] In one embodiment, searches for sets of multiple spellings may be performed, and one set of multiple spellings may be selected over another based on a comparison between a frequency of occurrences of a first set in the results of a search for the second set and a frequency of occurrences of the second set in the results of a search for the first set.

[0082] Figures 9A and 9B are flow diagrams that illustrate a technique, according to an embodiment of the present invention, for determining whether to retain or remove a candidate alternative spelling from a set of candidate alternative spellings based on how frequently the candidate alternative spelling occurs in result fields obtained from a search for another spelling. Through the technique illustrated below, a set of candidate alternative spellings may be filtered.

[0083] Referring to Figure 9A, in block 902, a search is performed for occurrences of a first spelling in a plurality of files. The first spelling may be, for example, a spelling that a user originally submitted in a search request. Results of the search for the first spelling are obtained as a first set of result fields.

[0084] In block 904, a search is performed for occurrences of a second spelling in the plurality of files. The second spelling may be, for example, a spelling that is similar to, but not exactly the same as, the first spelling. The second spelling may be a spelling that is returned by a spelling checking routine when the first spelling is submitted to the routine. The second spelling may be a spelling that occurs in anchor text of a page on which anchor text also contains the first spelling. Results of the search for the second spelling are obtained as a second set of result fields.

[0085] In block 906, it is determined whether a frequency of occurrences of the second spelling in the first set of result fields is greater than a frequency of occurrences of the first spelling in the second set of result fields. If the frequency of occurrences of the first spelling in the first set of result fields is greater than the frequency of occurrences of the first spelling in the second set of result fields, then control passes to block 908. Otherwise, control passes to block 910.

[0086] In block 908, the second spelling is retained in a set of candidate alternative spellings. The second spelling may ultimately be presented to a user as a suggested alternative spelling to the first spelling.

[0087] In block 910, it is determined whether a frequency of occurrences of the second spelling in the second set of result fields is greater than a frequency of occurrences of the first spelling in the first set of result fields. If the frequency of occurrences of the second spelling in the second set of result fields is greater than the frequency of occurrences of the first spelling in the first set of result fields, then control passes to block 908. Otherwise, control passes to block 912, which is illustrated in Figure 9B.

[0088] Referring to Figure 9B, in block 912, it is determined whether a frequency of occurrences of the first spelling in the first set of result fields is greater than a frequency of occurrences of the second spelling in the second set of result fields. If the frequency of occurrences of the first spelling in the first set of result fields is greater than the frequency of occurrences of the second spelling in the second set of result fields, then control passes to block 914. Otherwise, control passes to block 916.

[0089] In block 914, the second spelling is removed from a set of candidate alternative spellings. The second spelling will not be presented to a user as a suggested alternative spelling to the first spelling.

[0090] In block 916, it is not decided whether to retain or remove the second spelling from a set of candidate alternative spellings. That decision may be made, instead, based on some other criteria.

[0091] In one embodiment, if the difference between the average relevance score of the first set and the average relevance score of the second set is less than a specified number, then the second spelling is removed from a set of candidate alternative spellings.

[0092] In one embodiment, if a ratio of the frequency of the first spelling in the second set to the frequency of the second spelling in the first set is less than a specified number, then the second spelling is removed from a set of candidate alternative spellings.

[0093] In one embodiment, if the difference between the frequency of the second spelling in the second set and the frequency of the first spelling in the first set is greater than or equal to a specified number, then the second spelling is retained in a set of candidate alternative spellings.

[0094] In one embodiment, the second spelling is not retained in a set of candidate alternative spellings unless one or more specified conditions are satisfied.

GENERATING AND FILTERING CANDIDATE ALTERNATIVE SPELLINGS BY SEARCHING WHOLE SELECTED FILES FOR SIMILAR SPELLINGS

[0095] In response to a user entering a particular spelling into a search engine, the search engine selects, based on the particular spelling, one or more files from among a plurality of files. A file may be selected because the file contains the particular spelling. However, a file also may be selected because the file is "pointed to" or "linked to" by another file that contains the particular spelling in the text of the link to the selected file. Thus, a file may be selected regardless of whether the file contains the particular spelling entered by a user.

[0096] For example, if a user enters the spelling "Lincon" into a search engine, the search engine may find a first page that contains a link with the label "Lincon." The link may indicate the address of a second page, which may or may not contain the spelling "Lincon." Because the link to the second page contains the spelling entered by the user, the search engine may select the second page even if the second page does not contain the spelling entered by the user.

[0097] The second page may contain a spelling that is similar to, but not exactly the same as, the spelling entered by the user. For example, the second page may contain one or more instances of the spelling "Lincoln." Link text often contains more misspellings than non-link text, perhaps because authors of link text are sometimes less familiar with, or less concerned about, words used on pages authored by others. Therefore, the existence of the similar spelling on the second page is an indication that the spelling entered by the user and contained in the first page's link text might have been a misspelling. Therefore, the spelling on the second page is a good candidate for an alternative spelling.

[0098] Thus, according to one embodiment, in response to a user entering a first spelling, a search engine selects, from a plurality of pages, a page that contains a second spelling that is similar to the first spelling. The page may be selected because the page contains the first spelling, or because the page is linked to by a link that contains the first spelling. The search engine searches the page for spellings that are similar to the first spelling. In searching the page for similar spellings, the search engine may search the whole page, rather than only those portions of the page that would be displayed as result fields (e.g., the page title, abstracts generated for the page, the page's URL, etc.). As a result, the search engine locates one or more instances of the second spelling. In response to locating the second spelling, the search engine adds the second spelling to a list of candidate alternative spellings.

[0099] The search engine may filter the list of candidate alternative spellings according to techniques described herein. For example, the list of candidate alternative spellings may be filtered based on a comparison of frequencies of occurrences of the first and second spellings. For another example, the list of candidate alternative spellings may be filtered based on whether the first spelling is a plural form of the second spelling or vice-versa. The techniques described above may be applied not only to spellings of single words, but also to spellings of whole multiple-word phrases that are entered by users.

HARDWARE OVERVIEW

[0100] Figure 10 is a block diagram that illustrates a computer system 1000 upon which an embodiment of the invention may be implemented. Computer system 1000 includes a bus 1002 or other communication mechanism for communicating information, and a processor 1004 coupled with bus 1002 for processing information. Computer system 1000 also includes a main memory 1006, such as a random access memory (RAM) or other dynamic storage device, coupled to bus 1002 for storing information and instructions to be executed by processor 1004. Main memory 1006 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 1004. Computer system 1000 further includes a read only memory (ROM) 1008 or other static storage device coupled to bus 802 for storing static information and instructions for processor 1004. A storage device 1010, such as a magnetic disk or optical disk, is provided and coupled to bus 1002 for storing information and instructions.

[0101] Computer system 1000 may be coupled via bus 1002 to a display 1012, such as a cathode ray tube (CRT), for displaying information to a computer user. An input device 1014, including alphanumeric and other keys, is coupled to bus 1002 for communicating information and command selections to processor 1004. Another type of user input device is cursor control 1016, such as a mouse, a trackball, or cursor direction keys for communicating direction information and command selections to processor 1004 and for controlling cursor movement on display 1012. This input device typically has two degrees of freedom in two axes, a first axis (e.g., x) and a second axis (e.g., y), that allows the device to specify positions in a plane.

[0102] The invention is related to the use of computer system 1000 for implementing the techniques described herein. According to one embodiment of the invention, those techniques are performed by computer system 1000 in response to processor 1004 executing one or more sequences of one or more instructions contained in main memory 1006. Such instructions may be read into main memory 1006 from another computer-readable medium, such as storage device 1010. Execution of the sequences of instructions contained in main memory 1006 causes processor 1004 to perform the process steps described herein. In alternative embodiments, hard-wired circuitry may be used in place of or in combination with software instructions to implement the invention. Thus, embodiments of the invention are not limited to any specific combination of hardware circuitry and software.

[0103] The term "computer-readable medium" as used herein refers to any medium that participates in providing instructions to processor 1004 for execution. Such a medium may take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Non-volatile media includes, for example, optical or magnetic disks, such as storage device 1010. Volatile media includes dynamic memory, such as main memory 1006. Transmission media includes coaxial cables, copper wire and fiber optics, including the wires that comprise bus 1002. Transmission media can also take the form of acoustic or light waves, such as those generated during radio-wave and infra-red data communications.

[0104] Common forms of computer-readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, or any other magnetic medium, a CD-ROM, any other optical medium, punchcards, papertape, any other physical medium with patterns of holes, a RAM, a PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave as described hereinafter, or any other medium from which a computer can read.

[0105] Various forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to processor 1004 for execution. For example, the instructions may initially be carried on a magnetic disk of a remote computer. The remote computer can load the instructions into its dynamic memory and send the instructions over a telephone line using a modem. A modem local to computer system 1000 can receive the data on the telephone line and use an infra-red transmitter to convert the data to an infra-red signal. An infra-red detector can receive the data carried in the infra-red signal and appropriate circuitry can place the data on bus 1002. Bus 1002 carries the data to main memory 1006, from which processor 1004 retrieves and executes the instructions. The instructions received by main memory 1006 may optionally be stored on storage device 1010 either before or after execution by processor 1004.

[0106] Computer system 1000 also includes a communication interface 1018 coupled to bus 1002. Communication interface 1018 provides a two-way data communication coupling to a network link 1020 that is connected to a local network 1022. For example, communication interface 1018 may be an integrated services digital network (ISDN) card or a modem to provide a data communication connection to a corresponding type of telephone line. As another example, communication interface 1018 may be a local area network (LAN) card to provide a data communication connection to a compatible LAN. Wireless links may also be implemented. In any such implementation, communication interface 1018 sends and receives electrical, electromagnetic or optical signals that carry digital data streams representing various types of information.

[0107] Network link 1020 typically provides data communication through one or more networks to other data devices. For example, network link 1020 may provide a connection through local network 1022 to a host computer 1024 or to data equipment operated by an Internet Service Provider (ISP) 1026. ISP 1026 in turn provides data communication services through the world wide packet data communication network now commonly referred to as the "Internet" 1028. Local network 1022 and Internet 1028 both use electrical, electromagnetic or optical signals that carry digital data streams. The signals through the various networks and the signals on network link 1020 and through communication interface 1018, which carry the digital data to and from computer system 1000, are exemplary forms of carrier waves transporting the information.

[0108] Computer system 1000 can send messages and receive data, including program code, through the network(s), network link 1020 and communication interface 1018. In the Internet example, a server 1030 might transmit a requested code for an application program through Internet 1028, ISP 1026, local network 1022 and communication interface 1018.

[0109] The received code may be executed by processor 1004 as it is received, and/or stored in storage device 1010, or other non-volatile storage for later execution. In this manner, computer system 1000 may obtain application code in the form of a carrier wave.

[0110] In the foregoing specification, embodiments of the invention have been described with reference to numerous specific details that may vary from implementation to implementation. Thus, the sole and exclusive indicator of what is the invention, and is intended by the applicants to be the invention, is the set of claims that issue from this application, in the specific form in which such claims issue, including any subsequent correction. Any definitions set forth herein for terms contained in such claims shall govern the meaning of such terms as used in the claims. Hence, no limitation, element, property, feature, advantage or attribute that is not expressly recited in a claim should limit the scope of such claim in any way. The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.